DOCUMENT RESUME

ED 388 720 TM 024 184

AUTHOR Cowell, William R.
TITLE A Procedure for Estimating the Conditional Standard Errors of Measurement for GRE General and Subject Tests. GRE Board Professional Report No. 87-03P.
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY Graduate Record Examinations Board, Princeton, N.J.
REPORT NO ETS-RR-91-25
PUB DATE Jun 91
NOTE 17p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Entrance Examinations; *Error of Measurement; *Estimation (Mathematics); Higher Education; *Scores; Selection
IDENTIFIERS *Graduate Record Examinations; *Rights and Formula Scoring

ABSTRACT

A series of computer programs was written for computing the conditional standard errors of measurement (CSEM) for both rights-scored and formula-scored tests based on a method suggested by F. M. Lord (1984), commonly known as Lord's Method IV or the compound binomial method. These programs estimate the CSEM for both raw and scaled scores, average results for two or more forms, and compute form-to-form difference statistics for pairs of forms. CSEM, averages, and differences were computed for the verbal, quantitative, and analytical raw and converted scores for 8 forms of the Graduate Record Examinations (GRE) General Test and for 2 forms each of 15 GRE Subject Tests. The Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education recommend that test publishers provide estimates of the standard error measurement at a number of widely spaced score levels. The CSEM data produced in this study have been made available to three programs that use GRE scores, along with other criteria, for awarding fellowships. These data also have been made available for use in GRE program publications and in correspondence. (Contains five tables and eight references.) (Author/SLD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED 388 720

GRE®

RESEARCH

# A Procedure for Estimating the Conditional Standard Errors of Measurement for GRE General and Subject Tests

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

William R. Cowell

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

June 1991

GRE Board Professional Report No. 87-03P
ETS Research Report 91-25

BEST COPY AVAILABLE

TM024184

Educational Testing Service • Princeton, New Jersey

# A Procedure for Estimating the Conditional Standard Errors of Measurement for GRE General and Subject Tests

William R. Cowell

GRE Board Report No. 87-03P

June 1991

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

Educational Testing Service, Princeton, N.J. 08541

The Graduate Record Examinations Board and Educational Testing Service are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

Graduate Record Examinations and Educational Testing Service are U.S. registered trademarks of Educational Testing Service; GRE, ETS, and the ETS logo design are registered in the U.S.A. and in many other countries.

Copyright © 1991 by Educational Testing Service. All rights reserved.

Abstract

A series of computer programs were written for computing the conditional standard errors of measurement (CSEM) for both rights-scored and formula-scored tests based on a method suggested in Lord (1984), commonly known as Lord's Method IV or the compound binomial method. These programs estimate the conditional standard errors of measurement for both raw and scaled scores, average results for two or more forms, and compute form-to-form difference statistics for pairs of forms.

Conditional standard errors of measurement, averages, and differences have been computed for the verbal, quantitative, and analytical raw and converted scores for eight forms of the GRE General Test and for two forms each of 15 GRE Subject Tests.

The Standards for Educational and Psychological Testing (Committee of AERA, APA, & NCME to Develop Standards, 1985) recommends that test publishers provide estimates of the standard error measurement at a number of widely spaced score levels. The CSEM data produced in this study have been made available to three programs which use GRE scores, along with other criteria, for awarding fellowships. These data also have been made available for use in GRE program publications and in correspondence.

## Introduction

The concept of "standard error of measurement" is described in detail by Livingston (1988) and by Feldt and Brennan (1989). Briefly, the difference between the actual score a person obtained on a test and the hypothetical "true score" that person would have obtained if the test had been a perfect measure of the ability tested is called the "error of measurement." In practice, neither the true score nor the error of measurement can be determined for an individual examinee, but it is possible to estimate the mean and standard deviation of the errors of measurement for a group of examinees. We would expect the mean of the errors of measurement for a group of examinees to be zero because the positive and negative values would tend to balance out. The standard error of measurement (SEM) is an estimate of the standard deviation of the errors of measurement for a group of examinees. It has been customary to report, for each of the GRE tests, the average SEM for the test analysis sample. The SEM is a function of the reliability of the test and the variance of the scores for the analysis sample.

The standard error of measurement typically reported in a test analysis report is an average of the SEMs at the various score levels represented in the analysis sample. The SEM, however, is not the same at all score levels. The SEM is smaller for scores near the extremes of the score scale and larger near the middle of the scale. The SEM at a specified score level or ability level is referred to as the conditional standard error of measurement (CSEM). Procedures for estimating the CSEM values for number-right scores are described in the following paragraphs. However, to be useful to score recipients and score users, these estimates must be transformed to the scale used in score reporting. When the raw-to-scale conversion is not linear, the CSEM is strongly affected by curvilinearity in the conversion function. The use of the average SEM in place of the conditional SEM may be somewhat misleading in certain cases. For example, in the selection process for highly competitive fellowship awards the average SEM may be considerably larger than the conditional SEM for scores in the region of the scale where decisions are being made. The Standards for Educational and Psychological Testing (Committee of AERA, APA, & NCME to Develop Standards, 1985) recommends that test publishers provide estimates of the SEM at a number of widely spaced score levels.

Lord (1984) outlines four methods for estimating the conditional standard errors of measurement for number-right scores:

Method I. Matched half tests (split halves)

Method II. Item response theory (IRT)

Method III. Randomly parallel forms (binomial)

Method IV. Matched parrallel forms (compound binomial)

All of these methods apply to rights-scored tests (such as the General Test) but require modification for use with formula-scored tests (such as the Subject Tests).

Method II, the IRT method, has perhaps the best theoretical foundation and provides the most information, but it requires relatively large sample sizes and requires that the test be unidimensional. Dorans (1984) describes the development of procedures for computing approximate IRT formula score and scaled score conditional standard errors of measurement.

Method I, the split-halves method, has a solid theoretical foundation and applies directly to formula-scored tests as well as rights-scored tests, but it requires very large samples. Kingston (1985) proposed a project that has been funded by Educational Testing Service to develop computer programs for applying the split-halves method with data smoothing techniques. The data smoothing procedures may reduced the sample size requirements to some extent.

Method III, the binomial method, is the simplest of the four methods. It requires no data other than the number of items in the test. It is, however, of no practical interest because it is appropriate only when test forms are randomly parallel. It systematically overestimates the SEMs of tests for which forms are matched, such as on item difficulty level.

Method IV, the compound binomial method, is appropriate when forms are matched and requires neither complicated calculations nor large samples. For Method IV, the Method III SEM is multiplied by a factor that can easily be computed using ordinary test analysis data, as described in the following section.

In this study, the compound binomial method was used to obtain estimates of the conditional SEMs for two forms of each of 15 GRE Subject Tests and for eight forms of the General Test. This study has produced conditional standard error of measurement data that have been made available to three programs that use GRE scores, along with other criteria, for awarding fellowships. The data also have been made available for use in GRE program publications and in correspondence.

Feldt, Steffen, and Gupta (1985) compared five methods for estimating conditional standard errors of measurement, including these four methods. The compound binomial method described in their article, however, is based on clusters of items, as described in Lord (1965), rather than the variance of proportion correct used in this study, described in Lord (1984).

## Method for Computing the SEM for Number-Right Scores

For Method III, test forms are assumed to be random samples of n items selected from a universe of acceptable test items. In that case, the squared standard error of measurement is simply $t(n - t)/n$ , for any true score, t. For a randomly chosen examinee, an unbiased estimator of this quantity is $x(n - x)/(n - 1)$, for an observed score, x.

For test forms that are constructed to the same content and statistical specifications, the Method III estimates are too large. For Method IV, it is assumed that the appropriate squared SEM for matched forms is the Method III squared SEM multiplied by (1 - K). Lord (1965) derives the following value of K (with $2k = nK$):

$$K = [n(n - 1)S_p^2] \; / \; [M_x(n - M_x) - S_x^2 - nS_p^2]$$

where $S_p^2$ is the variance across items of proportion correct and $M_x$ and $S_x^2$ are the mean and variance across examinees of the number-right scores.

The standard error of measurement corresponding to each scaled score (GRE reported score) is estimated by multiplying the CSEM of the corresponding raw score by the local linear approximation to the slope of the conversion function at that raw score. The local linear approximation method uses the slope of the secant through the two closest adjacent points as that estimate. If more than one raw score converts to the same scaled score, the mean of the CSEM values is used. If no raw scores convert to the scaled score value, the mean of the CSEM values at adjacent raw scores is substituted. If conversions do not extend to ends of the scale, those values are dropped from tables and summary statistics.

## Method for Computing the SEM for Formula-Scored Tests

The formula score, FS, for the GRE Subject Tests is the number right (R) minus one-quarter the number wrong (W); FS = R - .25(W). For examinees who answer every item, FS = R - .25(n - R) = 1.25(R) - .25(n). If everyone answers every item, the SEM for any formula score is 1.25 times the SEM for the number-right score. For the GRE Subject Tests, however, this procedure cannot be used because there are too many omits to be ignored. The following procedures were developed for estimating the conditional standard errors of measurement for the GRE Subject Tests.

A computer program was written to read examinee data from a scores tape and compute linear conversion parameters for converting the rights scores to scaled scores. The program prints a scatterplot of rights scores versus formula scores and tables of means, standard deviations, and intercorrelations for rights, formula, and GRE scaled scores. The correlations tend to be in the high .90s, mostly .98 or .99. Conversion parameters relating rights scores to formula scores are computed by setting the means and the standard deviations for these scores equal. The parameters for converting rights scores to scaled scores are computed by algebraic substitution into the equation for converting formula scores to scaled scores obtained from linear methods of test score equating.

A second program computes the CSEM for each rights score by Lord's Method IV using test analysis data (number of items, sum of rights, sum of rights squared, and the Kuder-Richardson-20 (KR-20) reliability coefficient). It then computes the distribution of converted scores using the rights-score-to-scaled-score conversion parameters; prints a table of raw scores, converted scores, and CSEM values; computes the root mean square of the CSEM for the rights scores; and computes the CSEM for each scaled score using the rights-scores-to-scaled-score parameters and linear interpolation between adjacent unrounded converted scores. It also computes the root mean square of the CSEM of the scaled scores. If there are missing scaled scores, the program will average the CSEM values from the adjacent scores, and if the CSEM cannot be determined for scores at the end of the scale, the program will flag these scores.

The average of the CSEM values tended to be somewhat lower than the SEM computed from the KR-20 reliability coefficient. Therefore, the program also computes and tabulates adjusted SEM values (ASEM), by multiplying each CSEM by a factor to set the mean ASEM equal to the average SEM computed from the KR-20 reliability. These adjustments tend to be relatively small, the adjusted values being about 5 to 9% higher than the CSEM values.

Another program computes and tabulates the differences and the averages of the pairs of ASEM values from two or more forms of a test (nullifying values for one form at scores for which data are not available for the other form). It then computes the root mean square of the ASEM values for each form and the average values, each weighted by the sum of frequencies for the two forms at each value for which data are available. It also computes the root mean square, mean, and standard deviation of the differences between the ASEM values weighted by the sum of the frequencies for the two forms.

The standard error of measurement corresponding to each scaled score (GRE reported score) is estimated by the same procedure as described for number-right scores.

## Results

Tables 1 - 5 present the conditional standard errors of measurement computed using Lord's Method IV with transformations to formula scores (for GRE Subject Tests) and to scaled scores (for both GRE Subject and General Tests) by the procedures described on pages 3 and 4. Table 1 gives CSEM values at 50-point intervals for the GRE General and the GRE Subject Tests over the range of values for which data are available. These are appropriate for use in interpreting individual scores. In comparing the scores for two individuals, the difference will be affected by errors of measurement in both scores. Therefore, the standard error of measurement for the difference is larger than that of individual scores (by a factor of about 1.4). Table 2 gives the standard errors of measurement for score differences. Because scores are reported as multiples of 10 and score differences are therefore multiples of 10, "^" marks are used to separate values into groups that round to the same multiple of 10. Examples are given in the footnotes.

In general, the values of the conditional standard errors of measurement follow the expected pattern; that is, they tend to be somewhat smaller near the ends of the score scale and larger near the middle of the scale.

For the GRE General Test, the average of the CSEM values, weighted by the frequencies at each score level, matched very closely the average SEM values computed using Kuder-Richardson Formula 20.

For the GRE Subject Tests, the average of the CSEM values, after the transformations from rights to formula to scaled scores, were lower (by about 5% - 9%) than the corresponding KR-20 estimates. To avoid underestimates of the CSEM values, an adjustment was applied to CSEM values to set the mean of the adjusted values equal the mean based on KR-20. The tables and summary statistics are based on these adjusted (ASEM) values.

For the GRE Subject Tests, tables showing the average of and difference between the ASEM values for the two forms of each test are presented in Tables 3, 4, and 5. The differences, in particular, may be specific to the two forms that were selected for this study. The sign (+ or -) of the difference depends on which of the two forms was analyzed first and is of no particular significance except to show ranges where the signs are the same and where the signs are different. Whereas the SEM values tend to be smaller near the ends of the scale, the form-to-form differences tend to be larger near the ends of the scale. The magnitude of the differences for some of the test suggests that the ASEM values probably should be averaged over all active forms instead of just two forms. Method IV produces symmetrical CSEM values for the number-right scores. Some of that symmetry is lost in the nonlinear conversions to formula scores (for the GRE Subject Tests) and to scaled scores (for the GRE General Tests).

TABLE 1

Graduate Record Examinations
General Test and Subject Tests

STANDARD ERRORS OF MEASUREMENT*

Individual Scores
at Score Level

| GRE General Test | 200 | 250 | 300 | 350 | 400 | 450 | 500 | 550 | 600 | 650 | 700 | 750 | 800 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Verbal | 28.00 | 33.24 | 33.32 | 26.18 | 32.01 | 32.61 | ^38.24 | 36.40 | 36.48^ | 33.59 | 33.12 | 27.83^ | 16.02 |
| Quantitative | 33.74 | ^38.13 | 39.37 | 40.73 | 41.74 | 42.13 | 41.33 | 40.44 | 39.30 | 36.92^ | 32.46 | 25.43^ | 12.49 |
| Analytical | 33.05 | ^41.05 | ^48.77 | 48.81 | 48.70 | 49.91 | 49.19 | 46.59 | ^43.63 | 40.36 | 36.76 | ^31.29 | ^22.85 |

Individual Scores
at Score Level

| GRE Subject Tests | 250 | 300 | 350 | 400 | 450 | 500 | 550 | 600 | 650 | 700 | 750 | 800 | 850 | 900 | 950 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Biology | 16.02 | 19.79 | 22.46 | 24.40 | ^25.77 | 26.65 | 27.11 | 27.16 | 26.79 | 26.00 | ^24.74 | 22.94 | 20.46 | 16.98 | ^11.65 |
| Chemistry | --- | --- | 8.91 | ^17.14 | 21.78 | 24.96 | ^27.21 | 28.75 | 29.69 | 30.08 | 29.95 | 29.29 | 28.05 | 26.16 | ^23.45 |
| Computer Science | --- | --- | --- | 12.23 | ^20.65 | ^25.24 | 28.14 | 29.87 | 30.64 | 30.52 | 29.50 | 27.47 | ^24.17 | 18.92 | --- |
| Economics | --- | --- | --- | 16.58 | 21.37 | 24.47 | ^26.53 | 27.79 | 28.36 | 28.28 | 27.55 | 26.09 | ^23.77 | 20.23 | ^14.35 |
| Education | 17.93 | 20.08 | 21.50 | 22.31 | 22.60 | 22.38 | 21.64 | 20.32 | 18.29 | 15.27 | ^10.45 | --- | --- | --- | --- |
| Engineering | --- | --- | 15.94 | 21.84 | ^25.77 | 28.58 | 30.56 | 31.88 | 32.62 | 32.81 | 32.47 | 31.57 | 30.07 | 27.87 | ^24.78 |
| Geology | --- | 11.83 | ^16.82 | 19.97 | 22.13 | 23.56 | 24.41 | 24.72 | 24.53 | 23.81 | 22.52 | 20.53 | 17.58 | ^12.95 | --- |
| History | --- | 11.12 | ^15.04 | 17.37 | 18.76 | 19.41 | 19.41 | 18.76 | 17.37 | 15.05 | ^11.22 | --- | --- | --- | --- |
| Literature | 12.31 | ^15.49 | 17.55 | 18.87 | 19.60 | 19.80 | 19.48 | 18.64 | 17.17 | ^14.91 | 11.39 | ^ 4.06 | --- | --- | --- |
| Mathematics | --- | --- | 16.49 | ^26.85 | 33.16 | ^37.54 | 40.64 | 42.74 | 43.97 | 44.41 | 44.08 | 42.96 | 40.99 | 38.00 | ^33.71 |
| Music | 10.80 | ^15.13 | 17.80 | 19.53 | 20.57 | 21.02 | 20.93 | 20.29 | 19.04 | 17.04 | ^13.99 | 8.89 | --- | --- | --- |
| Physics | --- | --- | 20.20 | ^27.07 | 31.70 | ^35.05 | 37.46 | 39.11 | 40.10 | 40.47 | 40.25 | 39.42 | 37.94 | 35.73 | ^32.64 |
| Political Science | 15.00 | 18.77 | 21.26 | 22.89 | 23.84 | 24.20 | 23.98 | 23.18 | 21.72 | 19.47 | 16.08 | ^10.51 | --- | --- | --- |
| Psychology | 13.61 | ^17.53 | 20.12 | 21.87 | 22.99 | 23.57 | 23.65 | 23.23 | 22.28 | 20.74 | 18.45 | 15.08 | ^ 9.61 | --- | --- |
| Sociology | 19.94 | 22.43 | 24.16 | ^25.31 | 25.94 | 26.12 | 25.84 | 25.09 | ^23.82 | 21.96 | 19.31 | 15.48 | ^ 8.66 | --- | --- |
| | 250 | 300 | 350 | 400 | 450 | 500 | 550 | 600 | 650 | 700 | 750 | 800 | 850 | 900 | 950 |

Score Level

*Standard errors of measurement (SEM) for individual scores are shown at selected true score levels over the range of scores for which data are available. For the General Test, the values listed are averages for eight recent forms of the test. For the Subject Tests, the values listed are the averages for two recent forms. For individual scores, each value is an estimate of the standard deviation of observed scores about the given true score.

Note: Because scores are reported as multiples of 10, "^" marks are used to separate values into groups that round to the same multiple of 10. For example, the standard errors of measurement for individual scores on the GRE Subject Test in Biology at the 450, 500, 550, 600, 650, and 700 levels all round to 30 points and those at the 750, 800, 850, and 900 levels round to 20 points. At the 950 level the standard error of measurement is about 10 points.

TABLE 2

Graduate Record Examinations
General Test and Subject Tests

STANDARD ERRORS OF MEASUREMENT*

**Score Differences**
at Score Level

| GRE General Test | 200 | 250 | 300 | 350 | 400 | 450 | 500 | 550 | 600 | 650 | 700 | 750 | 800 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Verbal | 39.59 | 47.01 | 47.12 | 37.02 | 45.26 | 46.11 | 54.07 | 51.48 | 51.59 | 47.50 | 46.84 | 39.36 | 22.66 |
| Quantitative | 47.72 | 53.92 | 55.68 | 57.60 | 59.03 | 59.58 | 58.45 | 57.19 | 55.58 | 52.21 | 45.91 | 35.96 | 17.67 |
| Analytical | 46.74 | 58.05 | 68.97 | 69.03 | 68.87 | 70.58 | 69.56 | 65.89 | 61.70 | 57.07 | 51.99 | 44.25 | 32.31 |

**Score Differences**
at Score Level

| GRE Subject Tests | 250 | 300 | 350 | 400 | 450 | 500 | 550 | 600 | 650 | 700 | 750 | 800 | 850 | 900 | 950 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Biology | 22.66 | 27.99 | 31.77 | 34.51 | 36.44 | 37.70 | 38.34 | 38.40 | 37.89 | 36.77 | 34.99 | 32.45 | 28.93 | 24.02 | 16.48 |
| Chemistry | --- | --- | 12.61 | 24.24 | 30.80 | 35.30 | 38.49 | 40.66 | 41.99 | 42.54 | 42.35 | 41.42 | 39.67 | 36.99 | 33.16 |
| Computer Science | --- | --- | --- | 17.29 | 29.20 | 35.69 | 39.80 | 42.25 | 43.34 | 43.17 | 41.72 | 38.85 | 34.19 | 26.75 | --- |
| Economics | --- | --- | --- | 23.44 | 30.22 | 34.61 | 37.52 | 39.30 | 40.11 | 40.00 | 38.96 | 36.90 | 33.62 | 28.62 | 20.29 |
| Education | 25.36 | 28.40 | 30.40 | 31.56 | 31.97 | 31.66 | 30.61 | 28.73 | 25.86 | 21.60 | 14.77 | --- | --- | --- | --- |
| Engineering | --- | --- | 22.54 | 30.88 | 36.45 | 40.41 | 43.22 | 45.09 | 46.13 | 46.40 | 45.92 | 44.65 | 42.53 | 39.42 | 35.04 |
| Geology | --- | 16.73 | 23.79 | 28.25 | 31.29 | 33.32 | 34.52 | 34.97 | 34.69 | 33.67 | 31.84 | 29.03 | 24.87 | 18.32 | --- |
| History | --- | 15.72 | 21.27 | 24.56 | 26.52 | 27.45 | 27.45 | 26.53 | 24.57 | 21.29 | 15.86 | --- | --- | --- | --- |
| Literature | 17.41 | 21.90 | 24.82 | 26.69 | 27.71 | 28.00 | 27.55 | 26.35 | 24.28 | 21.08 | 16.10 | 5.74 | --- | --- | --- |
| Mathematics | --- | --- | 23.32 | 37.97 | 46.89 | 53.09 | 57.47 | 60.44 | 62.18 | 62.80 | 62.34 | 60.76 | 57.97 | 53.75 | 47.67 |
| Music | 15.27 | 21.40 | 25.17 | 27.61 | 29.08 | 29.73 | 29.60 | 28.69 | 26.92 | 24.10 | 19.78 | 12.58 | --- | --- | --- |
| Physics | --- | --- | 28.56 | 38.28 | 44.83 | 49.56 | 52.97 | 55.31 | 56.70 | 57.23 | 56.92 | 55.74 | 53.65 | 50.53 | 46.16 |
| Political Science | 21.22 | 26.54 | 30.06 | 32.37 | 33.72 | 34.22 | 33.91 | 32.78 | 30.72 | 27.53 | 22.74 | 14.86 | --- | --- | --- |
| Psychology | 19.25 | 24.79 | 28.45 | 30.93 | 32.52 | 33.34 | 33.44 | 32.85 | 31.51 | 29.33 | 26.09 | 21.33 | 13.59 | --- | --- |
| Sociology | 28.20 | 31.72 | 34.17 | 35.79 | 36.69 | 36.94 | 36.54 | 35.48 | 33.69 | 31.05 | 27.31 | 21.89 | 12.25 | --- | --- |
| | 250 | 300 | 350 | 400 | 450 | 500 | 550 | 600 | 650 | 700 | 750 | 800 | 850 | 900 | 950 |

Score Level

*Standard errors of Measurement (SEM) for score differences are shown at selected true score levels over the range of scores for which data are available. For the General Test, the values listed are averages for eight recent forms of the test. For the Subject Tests, the values listed are the averages for two recent forms. For score differences, each value is an estimate of the standard deviation of the differences between observed scores for two individuals, both at the same given true score level.

Note: Because scores are reported as multiples of 10 and score differences are therefore multiples of 10, "^" marks are used to separate values into groups that round to the same multiple of 10. For example, the standard errors of measurement for score differences on the GRE Subject Test in Biology at the 450, 500, 550, 600, 650, and 700 levels all round to 40 points and those at the 750, 800, and 850 levels round to 30 points. At the 900 and 950 levels the standard error of measurement of score differences is about 20 points.

TABLE 3

Averages and Differences of Standard Errors of Measurement
for Two Forms of 5 GRE Subject Tests at Selected Score Levels

| | Biology | | Chemistry | | Computer Science | | Economics | | Education | |
|---|---|---|---|---|---|---|---|---|---|---|
| C | AVER | DIF | AVER | DIF | AVER | DIF | AVER | DIF | AVER | DIF |
| *** | ***** | ***** | ***** | ***** | ***** | ***** | ***** | ***** | ***** | ***** |
| 950 | 12 | 2.2 | 23 | -4.8 | * | * | 14 | -13.5 | * | * |
| 900 | 17 | 1.5 | 26 | -3.9 | 19 | 6.7 | 20 | -8.9 | * | * |
| 850 | 20 | 1.2 | 28 | -3.3 | 24 | 4.7 | 24 | -6.8 | * | * |
| 800 | 23 | 1.1 | 29 | -2.9 | 27 | 3.5 | 26 | -5.3 | * | * |
| 750 | 25 | 1.0 | 30 | -2.6 | 30 | 2.7 | 28 | -4.2 | 10 | -.7 |
| 700 | 26 | .9 | 30 | -2.3 | 31 | 1.9 | 28 | -3.3 | 15 | -.8 |
| 650 | 27 | .9 | 30 | -2.2 | 31 | 1.2 | 28 | -2.4 | 18 | -.9 |
| 600 | 27 | .9 | 29 | -2.0 | 30 | .4 | 28 | -1.4 | 20 | -1.0 |
| 550 | 27 | .9 | 27 | -2.0 | 28 | -.5 | 27 | -.5 | 22 | -1.1 |
| 500 | 27 | .9 | 25 | -2.0 | 25 | -1.7 | 24 | .7 | 22 | -1.2 |
| 450 | 26 | 1.0 | 22 | -2.1 | 21 | -3.6 | 21 | 2.3 | 23 | -1.4 |
| 400 | 24 | 1.0 | 17 | -2.6 | 12 | -8.7 | 17 | 4.9 | 22 | -1.5 |
| 350 | 22 | 1.2 | 9 | -4.7 | * | * | * | * | 21 | -1.6 |
| 300 | 20 | 1.3 | * | * | * | * | * | * | 20 | -1.8 |
| 250 | 16 | 1.7 | * | * | * | * | * | * | 18 | -2.1 |
| 200 | 10 | 2.8 | * | * | * | * | * | * | 15 | -2.6 |
| RMS | 26 | 1.4 | 28 | 3.1 | 28 | 4.1 | 27 | 5.7 | 22 | 1.4 |

Notes: The conditional standard error of measurement (CSEM) at each number-right score was estimated using Lord's Method IV. Then parameters for converting rights scores to the GRE scale were used to estimate the CSEM at each scaled score. The adjusted standard errors of measurement (ASEM) are the CSEM values, each multiplied by a constant to set the mean ASEM equal to the mean SEM computed from the KR-20 reliability.

"AVER" is the average ASEM for two forms of each test.
"DIF" is the difference of the ASEM values for the two forms.

TABLE 4

Averages and Differences of Standard Errors of Measurement
for Two Forms of 5 GRE Subject Tests at Selected Score Levals

| | Engineering | | Geology | | History | | Literature in English | | Mathematics | |
|---|---|---|---|---|---|---|---|---|---|---|
| C | AVER | DIF | AVER | DIF | AVER | DIF | AVER | DIF | AVER | DIF |
| *** | ***** | ***** | ***** | ***** | ***** | ***** | ***** | ***** | ***** | ***** |
| 950 | 25 | 4.1 | * | * | * | * | * | * | 34 | 12.8 |
| 900 | 28 | 3.5 | 13 | 8.1 | * | * | * | * | 38 | 10.8 |
| 850 | 30 | 3.1 | 18 | 5.3 | * | * | * | * | 41 | 9.5 |
| 800 | 32 | 2.8 | 21 | 3.9 | * | * | 4 | -1.3 | 43 | 8.6 |
| 750 | 32 | 2.6 | 23 | 2.9 | 11 | -4.3 | 11 | -.1 | 44 | 7.8 |
| 700 | 33 | 2.4 | 24 | 2.2 | 15 | -3.1 | 15 | .1 | 44 | 7.2 |
| 650 | 33 | 2.3 | 25 | 1.6 | 17 | -2.6 | 17 | .2 | 44 | 6.6 |
| 600 | 32 | 2.2 | 25 | 1.1 | 19 | -2.4 | 19 | .3 | 43 | 6.2 |
| 550 | 31 | 2.1 | 24 | .6 | 19 | -2.4 | 19 | .3 | 41 | 5.8 |
| 500 | 29 | 2.1 | 24 | .1 | 19 | -2.4 | 20 | .3 | 38 | 5.4 |
| 450 | 26 | 2.1 | 22 | -.4 | 19 | -2.6 | 20 | .3 | 33 | 5.2 |
| 400 | 22 | 2.2 | 20 | -1.1 | 17 | -2.9 | 19 | .2 | 27 | 5.2 |
| 350 | 16 | 2.7 | 17 | -1.9 | 15 | -3.5 | 18 | .0 | 16 | 6.4 |
| 300 | * | * | 12 | -3.7 | 11 | -5.1 | 15 | -.2 | * | * |
| 250 | * | * | * | * | * | * | 12 | -.6 | * | * |
| 200 | * | * | * | * | * | * | 7 | -1.9 | * | * |
| RMS | 31 | 2.8 | 24 | 3.4 | 19 | 3.5 | 19 | 0.5 | 41 | 8.2 |

Notes: The conditional standard error of measurement (CSEM) at each number-right score was estimated using Lord's Method IV. Then parameters for converting rights scores to the GRE scale were used to estimate the CSEM at each scaled score. The adjusted standard errors of measurement (ASEM) are the CSEM values, each multiplied by a constant to set the mean ASEM equal to the mean SEM computed from the KR-20 reliability.

"AVER" is the average ASEM for two forms of each test.
"DIF" is the difference of the ASEM values for the two forms.

TABLE 5

Averages and Differences of Standard Errors of Measurement
for Two Forms of 5 GRE Subject Tests at Selected Score Levels

| | Music | | Physics | | Political Science | | Psychology | | Sociology | |
|---|---|---|---|---|---|---|---|---|---|---|
| C | AVER | DIF | AVER | DIF | AVER | DIF | AVER | DIF | AVER | DIF |
| *** | ***** | ***** | ***** | ***** | ***** | ***** | ***** | ***** | ***** | ***** |
| 950 | * | * | 33 | -4.5 | * | * | * | * | * | * |
| 900 | * | * | 36 | -3.6 | * | * | * | * | * | * |
| 850 | * | * | 38 | -2.8 | * | * | 10 | 2.7 | 9 | 7.4 |
| 800 | 9 | 3.2 | 39 | -2.2 | 11 | 1.9 | 15 | 2.5 | 15 | 4.4 |
| 750 | 14 | 2.3 | 40 | -1.6 | 16 | 1.4 | 18 | 2.6 | 19 | 3.7 |
| 700 | 17 | 2.1 | 40 | -1.1 | 19 | 1.2 | 21 | 2.7 | 22 | 3.5 |
| 650 | 19 | 1.9 | 40 | -.5 | 22 | 1.2 | 22 | 2.8 | 24 | 3.4 |
| 600 | 20 | 1.9 | 39 | .0 | 23 | 1.2 | 23 | 2.9 | 25 | 3.4 |
| 550 | 21 | 2.0 | 37 | .7 | 24 | 1.2 | 24 | 3.0 | 26 | 3.6 |
| 500 | 21 | 2.0 | 35 | 1.4 | 24 | 1.2 | 24 | 3.1 | 26 | 3.8 |
| 450 | 21 | 2.2 | 32 | 2.4 | 24 | 1.2 | 23 | 3.3 | 26 | 4.1 |
| 400 | 20 | 2.4 | 27 | 3.8 | 23 | 1.2 | 22 | 3.5 | 25 | 4.4 |
| 350 | 18 | 2.7 | 20 | 6.4 | 21 | 1.3 | 20 | 3.7 | 24 | 5.0 |
| 300 | 15 | 3.3 | * | * | 19 | 1.4 | 18 | 4.1 | 22 | 5.7 |
| 250 | 11 | 4.7 | * | * | 15 | 1.7 | 14 | 5.1 | 20 | 6.8 |
| 200 | * | * | * | * | 8 | 3.0 | * | * | 16 | 8 9 |
| RMS | 20 | 2.7 | 38 | 3.4 | 23 | 1.5 | 23 | 3.5 | 25 | 4.8 |

Notes: The conditional standard error of measurement (CSEM) at each number-right score was estimated using Lord's Method IV. Then parameters for converting rights scores to the GRE scale were used to estimate the CSEM at each scaled score. The adjusted standard errors of measurement (ASEM) are the CSEM values, each multiplied by a constant to set the mean ASEM equal to the mean SEM computed from the KR-20 reliability.

"AVER" is the average ASEM for two forms of each test.
"DIF" is the difference of the ASEM values for the two forms.

References

Committee of AERA, APA, & NCME to Develop Standards. (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Dorans, N. J. (1984). Approximate IRT formula score and scaled score standard errors of measurement at different score levels (Statistical Report SR-84-118), Princeton, NJ: Educational Testing Service.

Feldt, L. S., and Brennan, R. L. (1989) Reliability. In R. L. Linn (Ed.), Educational Measurement (3rd Ed.) New York: American Council on Education, Macmillan.

Feldt, L. S., Steffen, M., and Gupta, N. C. (1985). A comparison of five methods for estimating the standard error of measurement at specific score levels. Applied Psychological Measurement, 9, 351-361.

Kingston, N. M. (1985, February). Development of an appropriate method for estimating the overall and conditional standard errors of measurement for formula- and rights-scored test. Internal ETS memorandum to J. Murphy, Office of Corporate Quality Assurance.

Livingston, S. A. (1988). Reliability of test results. In J. P. Keeves (Ed.), Educational Research, Methodology, and Measurement, An International Handbook. Oxford: Pergamon Press.

Lord, F. M. (1965). A strong true-score theory, with applications. Psychometrika, 30, 239-270.

Lord, F. M. (1984). Standard errors of measurement at different score levels (Research Report RR-84-8). Princeton, NJ: Educational Testing Service.

54020-06555 • 561M 7 • 2098068